**Malnutrition estimates across the globe**

Description of variables

* Data of countries from 1983-2019.
* LDC - Least Developed Countries
* LIFD – Low Income Food Deficient
* LLDC or SIDS - Land Locked Developing Countries (LLDC) are 1, Small Island Developing States (SIDS) are 2, Others are 0
* Income Classification - Low Income: 0, Lower Middle Income: 1, Upper Middle Income: 2, High Income: 3
* Severe Wasting - % of children aged 0–59 months who are below minus three standard deviations from median weight-for-height
* Wasting – Moderate and severe: % of children aged 0–59 months who are below minus two standard deviations from median weight-for-height
* Overweight – Moderate and severe: % aged 0-59 months who are above two standard deviations from median weight-for-height
* Stunting – Moderate and severe: % of children aged 0–59 months who are below minus two standard deviations from median height-for-age
* Underweight – Moderate and severe: % of children aged 0–59 months who are below minus two standard deviations from median weight-for-age

1. **Part 1: Data Cleaning** 
   1. Load the dataset into R, ensure the fields are machine readable and check the structure of the dataset.
   2. Identify the columns with missing values. How many missing values are in the severe wasting column?
   3. Impute missing values in the following columns **(stunting, wasting, overweight, underweight, severe wasting)** with their respective column means.
   4. Convert the following variables into factors**(income classification, LLDC or SIDS, LIFD, LDC)**
   5. Remove all non-essential variables (e.g “**report author”, “iso-code, notes”**, etc.)
   6. Save the resulting dataset as “**mal\_clean**”.
2. **Part 2: Data Analysis** 
   1. Create a correlation matrix between all **relevant** numeric variables in the dataset.  
      How would you describe the relationship between “**stunting**” and “**wasting”** based on the matrix**?**
   2. Conduct a linear regression analysis to determine the relationship between **"underweight**” and all other variables excluding country
      1. Is the model a good fit for predicting underweight prevalence? Why or why not?
      2. What is the estimated effect of “**overweight**” on “**underweight**”, holding other variables constant?
   3. Identify the **top 10 countries** with the highest "**overweight**" value accompanied by their respective years. What is the top country & its overweight value and year?
3. **Part 3: Data Visualization** 
   1. Create a scatter plot that shows the relationship between "**wasting**" and "**stunting**". Use color to represent the "**LLDC or SIDS**" variable and include a trend-line for each state. Which type of countries/states have the highest “**wasting**” vs “**stunting**” rates?
   2. Create a box plot that shows the distribution of "**overweight**" by "**Income Classification**". What do you understand from this visual?
   3. Create a histogram that shows the distribution of “**underweight** ".
      1. Is the distribution of "**Underweight** “skewed? If so, in which direction? how would you interpret this

**Saving your outputs**

Name the folder containing your output using this format (Surname\_Firstname\_Meriam\_Test\_2023).

Please zip and email back the following:

* The script used to clean and analyse the data – provide snippets if you used a non-programming software [save as **Data\_cleaning\_analysis**\_**script**]
* All output files in form of graphs and tables [Save inside **Outputs** –folder]
* A word document/report with the relevant analysis outputs